P-Age: Pexels Dataset for Robust Spatio-Temporal Apparent Age Classification

Abid Ali, Marisetty Ashish, Francois F Bremond

P-Age: Pexels Dataset for Robust Spatio-Temporal Apparent Age Classification

Abid Ali*
Inria, France
Université Cote d'Azur, France
name.surname@inria.fr

Ashish Marisetty*
IIIT Naya Raipur, India
mashish1305@gmail.com

Francois Bremond
Inria, France
Université Cote d'Azur, France
name.surname@inria.fr

Abstract

Age estimation is a challenging task that has numerous applications. In this paper, we propose a new direction for age classification that utilizes a video-based model to address challenges such as occlusions, low-resolution, and lighting conditions. To address these challenges, we propose AgeFormer which utilizes spatio-temporal information on the dynamics of the entire body dominating face-based methods for age classification. Our novel two-stream architecture uses TimeSformer and EfficientNet as backbones, to effectively capture both facial and body dynamics information for efficient and accurate age estimation in videos. Furthermore, to fill the gap in predicting age in real-world situations from videos, we construct a video dataset called Pexels Age (P-Age) for age classification. The proposed method achieves superior results compared to existing face-based age estimation methods and is evaluated in situations where the face is highly occluded, blurred, or masked. The method is also cross-tested on a variety of challenging video datasets such as Charades, Smarthome, and Thumos-14. The code and dataset is available at https://github.com/Ashish013/AgeFormer.

1. Introduction

Age estimation is a challenging task that has many applications in various domains such as social robotics, video surveillance, business intelligence, social networks, and demography. Typically, the goal of age estimation involves predicting the age of a person by his/her facial appearance, which can be affected by many factors such as resolution, lighting condition, pose, expression, occlusion, and makeup. In recent years, numerous attempts have been made to estimate the age of a person from facial images [25, 27, 33, 35, 41, 51, 53, 55].

Age estimation models can be broadly classified into two categories: regression-based and classification-based. Regression-based models directly output a numerical value for the age, whereas classification-based models assign the image to one of the predefined age groups. Recent advances in age estimation are mainly driven by the development of deep learning techniques, especially Convolutional Neural Networks (CNNs). Several methods adopted 2D CNNs to estimate age from a single-face image. Furthermore, researchers recently combined 2D-CNNs with attention-mechanism [45], adversarial learning [31], domain adaptation [39], and multi-task learning techniques [20, 49] to improve the robustness of age estimation models. However, estimating the age of a single-face image is still an open challenge due to several factors such as **occlusions**, **low-resolution**, and **lighting conditions**. In addition to that, face-based age estimation models fail in situations where the face is not **visible** all the time or **blurred** due to privacy concerns. We argue that a single-face image is not enough to overcome these challenges in real-world situations. Therefore, a new technique based on video that captures the semantics of the entire body is needed to address these challenges.

3D-CNNs have been widely used in domains such as video analysis [15, 42], action recognition [4, 22, 47], and medical image segmentation [29], due to their ability to capture spatial and temporal information together. A video-based model can be useful to capture the spatio-temporal features of the frames to predict the age of a person in real-world situations such as surveillance and or medical assessment, where the face is not visible all the time. In addition, modern techniques, such as transformers [44] can be used to further handle long-range dependencies in an environment where the face is highly occluded. Furthermore, we believe that using only facial features is not enough to predict age. In addition, other characteristics of the body, such as the **body dynamics of a person** (face, head, and limbs, etc.) can provide significant cues to estimate the age of the person. Therefore, we propose a video-based age estimation model that captures significant spatio-temporal features of the entire body to improve the age estimation task. We reuse existing video models for a completely new problem (age

* authors contributed equally

predictions). Our results on unseen data such as Charades, Smarthome, and Thumos-14 Figure 5 demonstrate the novelty of utilizing such video models for a completely different problem such as age classification. This opens up new research directions for problems other than age classification, such as gender, ethnic groups, and other demographic classifications. Many works that rely on faces can be extended to video-based models to capture body dynamics for robust recognition.

In summary, in this paper, we propose a new direction for age estimation utilizing video-based models to address the above challenges. We introduced a new dataset, **P-Age**, for video-based age classification. Furthermore, we propose a novel two-stream architecture to improve age prediction in real-world scenarios. Our architecture uses TimeSformer [6] in combination with EfficientNet [40] as the backbone. The features of both streams are fused using a multihead-attention module for age classification. Some of the main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to propose a video-based method for age classification in challenging situations. Our novel **AgeFormer** architecture achieves superior results compared to existing face-based age estimation methods.
- We propose a new dataset called Pexels Age (**P-Age**), scrapped from Pexels (a royalty-free stock footage website) for age classification into four distinct groups. Additionally, we provide baseline results on this dataset and compare our method with existing face-based methods on the P-Age-Face dataset.
- Keeping privacy preservation a priority, we evaluate our method in situations where the face is i) **extremely occluded**, ii) **blurred**, and iii) **masked**. In addition, we validate the efficiency of our method in real-world situations by testing it on a variety of challenging video datasets such as *Charades* [37], *Smarthome* [12], and *Thumos-14* [3].

2. Related Work

Age Estimation: Deep learning-based methods use convolutional neural networks (CNNs) to automatically learn features from large-scale face datasets [25, 28, 34, 36]. Age-pattern learning methods can be categorized into classification-based and regression-based approaches. Classification-based methods treat age estimation as a multi-class problem, where each class corresponds to an age group or a single year [18, 24, 25]. AGNet [27] uses CNNs network with a softmax layer to classify the images into 8 age groups. Recently, PML [13] proposes a progressive marginal loss approach for the classification of unconstrained facial age.

Regression-based methods treat age estimation as a continuous problem, where the output is a real-valued age estimate. DEX [33] uses CNNs to learn the mapping of image pixels to age labels and then computes the expected value of the age distribution as the final prediction. OR-CNN [28] converts the age estimation problem into an ordinal regression problem, where each neuron in the output layer represents an ordinal number, and the model learns to activate neurons that correspond to the true age or lower. SSR-Net [48] adopts a multi-stage design that progressively refines age prediction from coarse to fine. More recently, [36] proposed a new ordinal regression algorithm called Moving Window Regression (MWR) that introduces the notion of relative rank (rho-rank) and develops global and local relative regressors (rho-regressors) to predict rho-ranks within entire and specific rank ranges, respectively, for age estimation from a single face image.

Furthermore, several datasets have been proposed for age estimation in recent years, including UTKFace [8], MORPH [32], CACD [9], FG-NET [30]. However, these datasets are image-based mainly focusing on face-crops. Therefore, the methods introduced in the past are limited to face-crops only. In contrast, we propose a video-based model to learn spatial and temporal features from the semantics of the entire body.

Video Classification: CNNs have shown great success in learning 3D spatio-temporal representations to recognize human actions [7, 19, 42, 43]. Two-stream techniques, frequently combined RGB with optical flow [17, 38], to classify videos. SlowFast network [16] has shown that action recognition can be enhanced by mixing representations of various temporal resolutions (i.e. frame rates). More recently, with the introduction of Transformers, numerous systems have improved action recognition by including attention [5, 10, 11, 26, 46]. TimeSformer [6] introduced a divided space-time attention mechanism to capture spatio-temporal dependencies throughout the video.

To go beyond Video Classification, our architecture utilizes action recognition architectures for age estimation tasks. We combine an action recognition (video) model with a 2D-CNNs method (image) to create a two-stream network for robust age classification.

3. P-Age Dataset

In this section, we discuss the key choices in creating the Pexels Age (P-Age) classification dataset. First, we discuss the data scraping process. Then, we explain the annotation procedure and the quality control checks to refine the annotations.

3.1. Data Preparation

The P-Age dataset is sourced from Pexels (a royalty-free stock footage website) [2] using search keywords such as

Figure 1. An overview of the P-Age dataset.

Age Group	Downloaded Videos	Filtered Videos	# of Frames		
			Min.	Avg.	Max.
Baby / Toddler	1760	401	4	315	1708
Adolescent	880	531	2	393	2436
Adult	1048	416	18	485	3523
Elderly	880	324	2	426	2106
Total	4568	1672	-	-	-

Table 1. Statistics of the P-Age dataset.

baby, toddler, teen, man, woman, elderly, old, etc.. Furthermore, our dataset represents different ethnic groups (*African, European, etc.*). An overview of the dataset is given in Figure 1. We downloaded more than 4,500 videos with 27-fps and a high-quality 1080p resolution. The average video has more than 400 frames and 5.7% of the videos in the P-Age dataset do not have a face appearance. Furthermore, most of the videos in our dataset have occluded faces (for example, a cinematic video of a child playing initially from his/her legs and slowly changing the camera to reveal the face).

3.2. Annotation

The P-Age dataset is divided into four age groups: baby/child, teen, adult, and elderly based on the searched query. For each class, we downloaded at least 880 videos per class and made sure that the gender ratio was balanced by using gender-specific keywords such as *baby-girl, teen-boy, etc.*. To further save time, we annotated the dataset based on the searched query type (child, teen, adult, elderly).

3.3. Quality Control

The downloaded videos are then cleaned with their brief video filenames to ensure that the video actually belongs to the class to which it is assigned, and videos that do not belong to any class are removed. Several videos have more than one person. We use existing object detection and tracking tools to filter out videos that have multiple people or no people. We utilize Yolov5 [21] with OSNet [54] to fil-

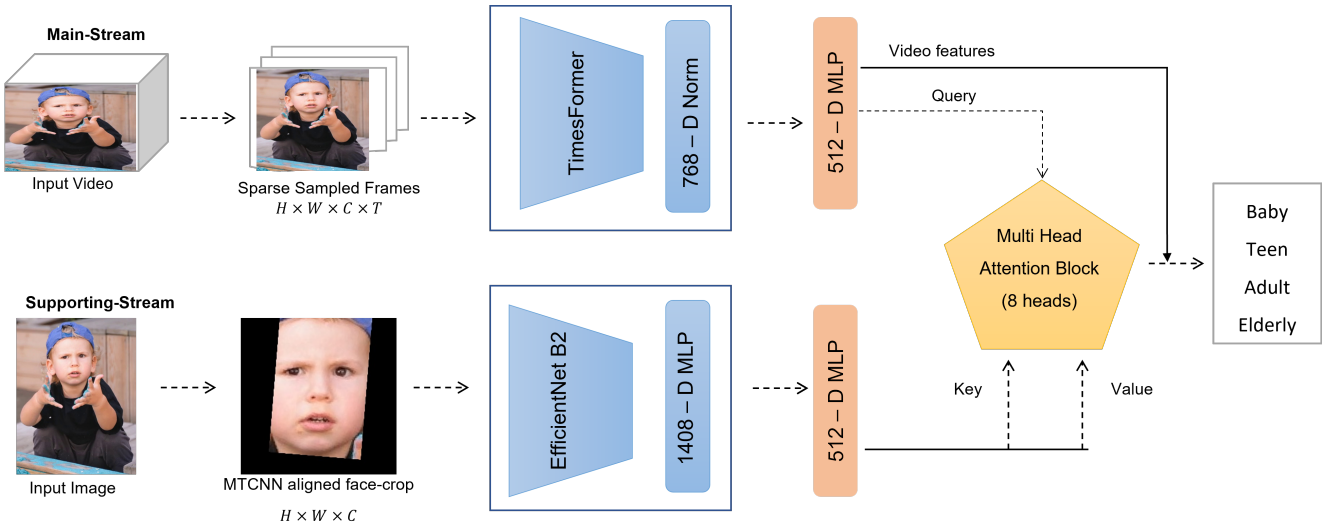

Figure 2. Proposed AgeFormer architecture. The main-stream learns spatio-temporal features of body dynamics, while the supporting-stream provides additional face features to improve age prediction.

	# videos	# images	occlusions	# videos having no faces
MORPH [32]	×	55.1k	×	×
UTKFace [8]	×	23.7k	×	×
P-Age	1672	675.5k	✓	95

Table 2. Comparison with age estimation benchmark datasets.

ter out such videos. Furthermore, three experts visually cross-check the annotation to correct any incorrect videos and remove outliers. After filtering, we have obtained 1672 videos; per-class statistics of the dataset are given in Table 1. A comparison is noted in Table 2 with existing age estimation datasets.

3.4. P-Age-Face

Furthermore, to compare with other age estimation methods, we have created a face-only variant of the P-Age dataset called **P-Age-Face** by randomly sampling 10 frames in a stride of 4 from each video. Following the general pre-processing convention outlined in [25], we extract the faces using MTCNN [52] and then use the eye landmark coordinates as a reference to align the face crops. Since 5.7% of the videos in our dataset lack faces, the P-Age-Face dataset accounts for 94. 3% of the size of the P-Age dataset. Upon acceptance, both datasets will be made available to the general public.

4. AgeFormer

In this section, we discuss our proposed architecture for robust age classification in videos. Our design is inspired by action recognition networks and how they handle the spatio-temporal information to recognize an action type. In contrast to action recognition methods, where the model learns gestures and body movements over time, we learn the dynamics of the entire body (*face, head, height and width of limbs, etc.*) over time. Here, the time information helps the model to learn the appearance of the person at different frames in a video snippet. For example, in the starting frames of a video, only the lower body is visible to the camera, but after a few seconds, the entire person appears (face or upper body appearance). Thus, the full appearance information is compatible with partially visible frames. The overall architecture is shown in Figure 2.

4.1. Video-stream

The main-stream is a video transformer network that uses TimeSformer [6] as the backbone. TimeSformer [6] is a convolution-free method to classify videos that relies solely on individual attention in space and time. By allowing spatio-temporal features to learn directly from a series of frame-level patches, it applies the traditional Transformer architecture to videos. The main-stream takes an input clip $X \in \mathbb{R}^{H \times W \times 3 \times T}$ having T frames of size $H \times W$. The

input frames are decomposed into N patches of size $P \times P$. The patches are flattened into vectors $X_{(p,t)}\mathbb{R}^{3P^2}$, where $p = 1, ..., N$ denotes spatial locations and $t = 1, ..., T$ defines the frame index. The readers can refer to the TimeSformer [6] paper for more details. The main-stream captures the spatio-temporal dynamics of the entire body.

4.2. Image-Stream

The supporting stream is a 2D-CNN network that uses EfficientNet B2 [40] as the backbone. EfficientNet is a robust and high-performing image classification network that has fewer parameters and is less computationally expensive, making it the perfect choice to complement the parameter-heavy video backbone. The supporting stream takes the face-crops extracted using MTCNN [52], aligned by the eye landmarks from the P-Age-Face dataset as input $X \in \mathbb{R}^{H \times W \times 3}$, where H and W represent height and width. This stream learns additional facial features that are known to be good indicators to further improve age estimation. In addition, a zero matrix is provided as input to the supporting stream when there is no face information available. More details of EfficientNet can be found in [41].

4.3. Attention Mechanism

A 768 feature vector from the main-stream, and a 1408 feature vector from supporting-stream are passed through an MLP layer obtaining 512 dimensional vectors for each stream after applying dropout and layer-normalization. The two streams are concatenated using a multi-head-attention (MHA) module having 8 attention heads. Query Q is obtained from the main-stream attended with the Key and Value $\{K, V\}$ pair acquired from the supporting-stream. A skip connection is added between the main-stream output and the MHA module output and passes through a linear layer with a softmax activation function to predict the age class. This architecture allows our model to effectively capture both facial and body dynamics information in the input for efficient and accurate age estimation using videos.

5. Experiments

Our experiments are categorized into four folds. First, we compare different video-based methods on the P-Age dataset, providing baseline results. Second, we compare our AgeFormer with existing face-based SOTA methods on the P-Age-Face dataset. Third, we evaluate the quality of our method in robust situations such as privacy preservation (occluded, blurred, or blacked-out faces). Finally, we cross-test AgeFormer on different challenging video datasets such as THUMOS-14 [3], Smarthome [12] and Charades [37]. Further details such as analysis of low-resolution, the effect of different sampling-rates, and situations where our model did not perform well are provided in the Supplementary Materials.

5.1. Experimental Details

All our models are implemented using the PyTorch library. We leverage the pre-trained weights from Kinetics-400 and ImageNet to initialize the models for the video and image streams, respectively. The main-stream receives 32 (temporal length) 224×224 resolution frames sampled at a stride of 4, while the supporting-stream takes in a 288×288 resolution face crop. The proposed architecture and video baselines are trained for 25 epochs with a batch size of 16 using an AdamW optimizer with a learning rate of $3e^{-5}$ and an exponential learning rate scheduler with a gamma of 0.9 using a cross-entropy loss. On the other hand, the face classifier benchmark training is conducted for 200 epochs, with a batch size of 512 and a learning rate of $2e^{-3}$. These hyper-parameter settings have been configured to ensure the convergence of all models. Furthermore, all of the performance analysis experiments involve only changing one parameter in the study (resolution, face-blur, stride) at a time, while the rest of the model remains unchanged and in inference mode.

Moreover, The P-Age dataset is split in a stratified way into train, validation, and test sets that have a proportion of 76%, 12%, and 12%, respectively.

Method	Acc.	Precision/Recall	F1-Score
X3D [15]	77.61%	0.80 / 0.76	0.77
Slow-Fast [16]	78.11 %	0.79 / 0.78	0.78
I3D [7]	81.59%	0.82 / 0.82	0.82
MViT [14]	86.00%	0.86 / 0.87	0.86
TimeSformer [6]	87.56%	0.88 / 0.88	0.88
Proposed	**89.55%**	**0.90 / 0.90**	**0.90**

Table 3. Baseline results of different video-based methods on the P-Age dataset.

5.2. Results and Discussion

We divide our experiments into two categories. First, we create a baseline for video-based age classification. We evaluate different action classification architectures for the age-estimation task on the P-Age dataset. We compare both 3D-CNNs and transformer-based methods for this task, as shown in Table 3. Our AgeFormer achieves SOTA results. Furthermore, the results indicate that video-based methods, specifically transformer architectures, are good enough to estimate the age of an individual. Adding an additional face-branch as in our AgeFormer, can further improve accuracy.

Second, we utilize the Face variant of the P-Age dataset to make use of the SOTA face-based methods for comparison with our AgeFormer. AgeFormer achieves new SOTA results compared to face-based methods on the P-

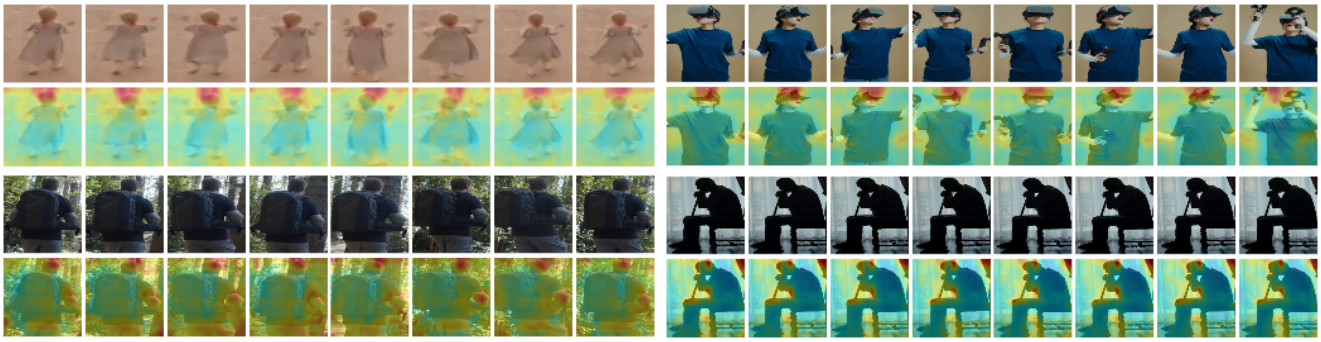

Figure 3. Spatio-temporal attention visualization of AgeFormer on P-Age dataset.

Figure 4. Comparison of SOTA face-based age classifier [25] with our proposed video-based method. Our method takes a video input with blurred faces.

Method	Acc.	Precision/Recall	F1-Score
C3AE [50]	37.70%	0.21 / 0.30	0.21
MWR [36]	43.39%	0.40 / 0.42	0.39
Kim [23]	57.67 %	0.68 / 0.58	0.59
ResNet-18 [1]	61.38 %	0.68 / 0.60	0.61
Levi [25]	70.37%	0.70 / 0.69	0.69
EfficientNet B2	53.97%	0.66 / 0.55	0.52
Proposed*	**77.25%**	**0.79 / 0.77**	**0.77**
Proposed	**89.42%**	**0.89 / 0.90**	**0.89**

Table 4. Comparison of existing face-classifiers with proposed video-based classifiers on P-Age-Face dataset. * means a video-based model (only main-stream) that utilizes the whole body with blurred faces as input. EfficientNet B2 results are from our supporting-stream (face model).

Age dataset. We compare our method with face-based methods in two ways as noted in Table 4. The results indicate that our proposed method is superior to existing methods, even when the faces are blurred. This validates our idea of using space-time dependencies of the entire body. A more qualitative comparison between Ageformer and SOTA method is provided in the Supplementary Materials.

5.3. Important Cues for Predicting Age

We visualize the divided space-time attention weights of the video backbone to provide insight into which part of the input is the most important when predicting age. The body of a person (limbs, chest, torso) and its dynamics hold important cues to differentiate one age group from the other. Our model learns these cues as shown in Figure 3. Using the attention roll-out method, we can visualize that the model is attentive to not just the face, but also other body parts, especially in the case where the face is not visible Figure 3 (second row, first example). Furthermore, the prediction of age from the body dynamics of a person has been validated in unseen videos in Figure 5 (Smarthome), where the faces are barely visible, but our model can accurately predict age in such situations.

5.4. Privacy Preservation

In this section, we discuss the privacy preservation aspect of our method. In real-world applications such as activities of daily living, surveillance, and medical evaluations, most of the data are sensitive to privacy concerns, where

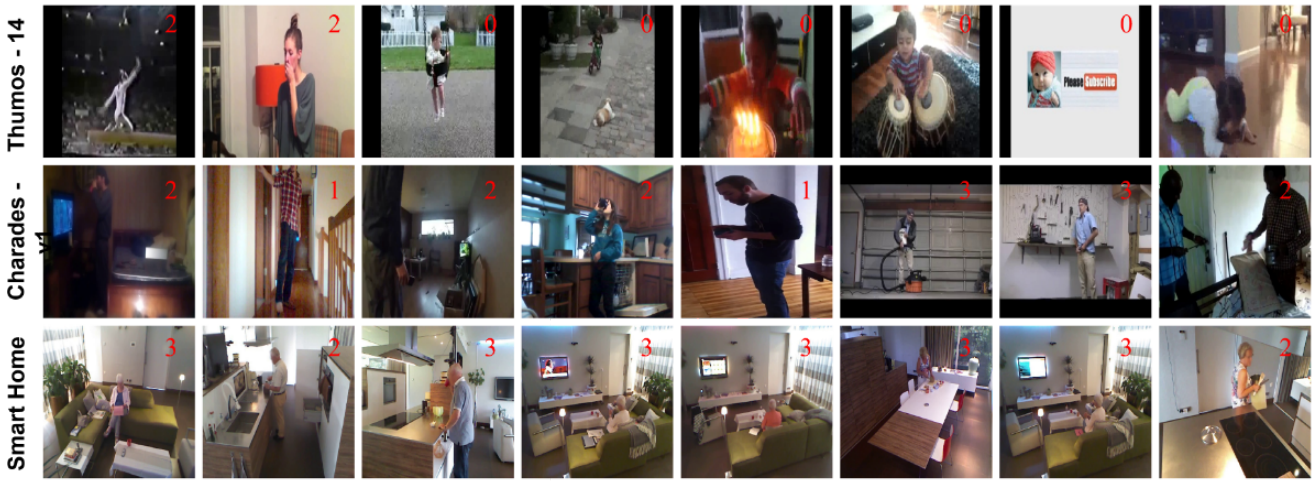

Figure 5. Each row depicts eight randomly selected videos from a specific dataset, the annotated label is predicted by AgeFormer. We observe that despite having never been trained on them, the model performs reasonably well.

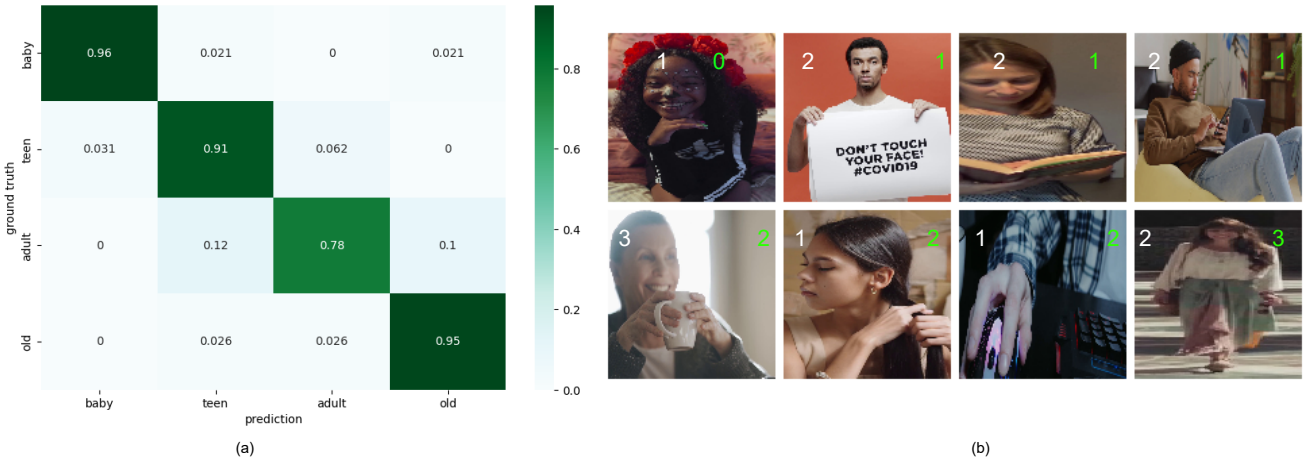

Figure 6. (a) illustrates the confusion matrix of the proposed method and (b) shows bad cases of our proposed method. White color indicates ground-truth, and green indicates the predicted label.

Augmentation	Acc.	Pre./Recall	F1-Score
face-blur	77.61 %	0.80 / 0.78	0.78
face blacked-out	76.88%	0.79 / 0.76	0.76

Table 5. Effectiveness of the proposed approach in the situation of extreme privacy preservation in the P-Age dataset.

faces are often obscured, blurred, or blacked-out. Furthermore, in cases such as autism assessment for children [4], the movements of the child cannot be controlled, and therefore their faces are not visible all the time, or, in other words, the upper-body is highly occluded. In such cases, existing face-based methods do not categorize the age of the participants.

Therefore, our AgeFormer is superior to prior face-based methods by incorporating spatio-temporal features of the entire body. This helps the model classify the age more efficiently, as demonstrated in Figure 4, even in scenarios where the faces are blurred-out. In these experiments, we use only the main-stream of our architecture. We further evaluate our model on two scenarios, i) blurred, and ii) blacked-out faces. The results in Table 5 validate our idea of using a video-based model with the whole body as input to preserve privacy.

5.5. Qualitative Cross Dataset Performance

In this section, we evaluate the generalizability of our proposed AgeFormer. We perform a qualitative cross-dataset performance analysis on real-world video (action recognition) benchmarks such as THUMOS-14 [3], Smarthome [12], and Charades [37] without training on these datasets. These datasets are selected because together they contain people from all age groups with significant variations in facial expressions, poses, lighting conditions, and occlusions. To this extent, we first randomly sample 30 different videos from each dataset and perform person detection and tracking to obtain the various person videos (tracklet). Following that, we use AgeFormer to predict the age class of each person (tracklet). As there are no ground-truth labels available for these datasets, we assign only the predicted labels to each person, as illustrated in Figure 5.

Our results show that, although the model was never trained on low-resolution data, the proposed approach is resilient in most scenarios, except for a few cases where the video quality is severely degraded (Figure 5 (1,1)), extreme low light conditions (Figure 5 (2,8)), or the apparent age of the person is uncertain and falls between two categories (Figure 5 (3,8)). If a facial crop is achievable, the latter issue could be potentially resolved by leveraging the facial feature context of the supporting stream, leading to more consistent predictions and better performance. In general, the proposed two-stream method is an effective approach with broad generalizability by extracting the relevant contextual features from the inputs.

5.6. Bad Cases

The proposed method may not perform well in tricky situations, where it is even difficult for humans to predict the age of an individual, as shown in Figure 6b. The confusion matrix in Figure 6a demonstrates that the model has a hard time classifying adults. In such difficult cases, the proposed model mostly confuses the teen and adult classes. Furthermore, the visual appearance of the individuals in the first column of Figure 6b illustrates why the model incorrectly predicts age.

6. Conclusion

Age-classification of an individual in real-world situations is a challenging task. In this paper, we propose a new direction to predict the age of an individual in a video. Our novel video-based model named AgeFormer achieves a precise age classification in challenging situations. The proposed architecture utilizes spatio-temporal information of the dynamics of the entire body dominating face-based methods for age classification. Experiments illustrate that video-based models are robust against situations where the face is obscured (privacy preservation) to predict the age of an individual. Additionally, we built the first video dataset (P-Age) for age classification, opening the door for researchers to explore video-based methods for age classification.

Acknowledgments

The COFUND BoostUrCAreer program has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Curie grant agreement No 847581, from the Région SUD Provence-Alpes-Côte d'Azur and IDEX UCAjedi.

This work is also supported by the French government, through the ACTIVIS project managed by the National Research Agency (ANR) with the reference number ANR-19-CE19-0004.

References

[1] Nebula,resnet-18 trained on afad dataset github repository. https://github.com/Nebula4869/PyTorch-gender-age-estimation. Accessed: 2023-05-01. 6

[2] Pexels, royalty-free stock footage website. https://www.pexels.com. Accessed: 2023-05-01. 2

[3] The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 2, 5, 8

[4] Abid Ali, Farhood F Negin, Francois F Bremond, and Susanne Thümmler. Video-based Behavior Understanding of Children for Objective Diagnosis of Autism. In *VISAPP 2022 - 17th International Conference on Computer Vision Theory and Applications*, Online, France, Feb. 2022. 1, 7

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021. 2

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 2, 4, 5

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 5

[8] Praveen Kumar Chandaliya, Vardhman Kumar, Mayank Harjani, and Neeta Nain. Scdae: Ethnicity and gender alteration on clf and utkface dataset. In *Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part II*, pages 294–306. Springer, 2020. 2, 4

[9] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Face recognition and retrieval using cross-age reference coding

with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015. 2

[10] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S. Ryoo, and François Brémond. Ms-tct: Multi-scale temporal convtransformer for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20041–20051, June 2022. 2

[11] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Pdan: Pyramid dilated attention network for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2970–2979, January 2021. 2

[12] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2533–2550, 2022. 2, 5, 8

[13] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. Pml: Progressive margin loss for long-tailed age classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10503–10512, June 2021. 2

[14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 5

[15] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 1, 5

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 5

[17] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 2

[18] Xin Geng, Zhi-Hua Zhou, Yu Zhang, Gang Li, and Honghua Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 307–316, 2006. 2

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017. 2

[20] Yoosoo Jeong, Seungmin Lee, Daejin Park, and Kil Houm Park. Accurate age estimation using multi-task siamese network-based deep metric learning for frontal face images. *Symmetry*, 10(9), 2018. 1

[21] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, (Zeng Yifu), Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Nov. 2022. 3

[22] Kumara Kahatapitiya and Michael S. Ryoo. Coarse-fine networks for temporal activity detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8385–8394, June 2021. 1

[23] Taewoon Kim. Generalizing mlps with dropouts, batch normalization, and skip connections. *arXiv preprint arXiv:2108.08186*, 2021. 6

[24] Young H Kwon and Niels da Vitoria Lobo. Age classification from facial images. *Computer vision and image understanding*, 74(1):1–21, 1999. 2

[25] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015. 1, 2, 4, 6

[26] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, June 2022. 2

[27] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24, 2015. 1, 2

[28] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016. 2

[29] S. Niyas, S.J. Pawan, M. Anand Kumar, and Jeny Rajan. Medical image segmentation with 3d convolutional neural networks: A survey. *Neurocomputing*, 493:397–413, 2022. 1

[30] Gabriel Panis, Andreas Lanitis, Nicholas Tsapatsoulis, and Timothy F Cootes. Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 5(2):37–46, 2016. 2

[31] Sun Penghui, Liu Hao, Wang Xin, Yu Zhenhua, and Suping Wu. Similarity-aware deep adversarial learning for facial age estimation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 260–265, 2019. 1

[32] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 341–345. IEEE, 2006. 2, 4

[33] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 1, 2

[34] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4030–4038, 2017. 2

[35] Vikas Sheoran, Shreyansh Joshi, and Tanisha R Bhayani. Age and gender prediction using deep cnns and transfer learning. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*, pages 293–304. Springer, 2021. 1

[36] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: a novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18760–18769, 2022. 2, 6

[37] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 2, 5, 8

[38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 2

[39] Ankita Singh and Shayok Chakraborty. *Deep Domain Adaptation for Regression*, pages 91–115. Springer International Publishing, Cham, 2020. 1

[40] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 2, 5

[41] Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan Z Li. Efficient group-n encoding and decoding for facial age estimation. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2610–2623, 2017. 1, 5

[42] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: generic features for video analysis. *CoRR, abs/1412.0767*, 2(7):8, 2014. 1, 2

[43] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[45] Haoyi Wang, Victor Sanchez, and Chang-Tsun Li. Improving face-based age estimation with attention-based dynamic patch fusion. *IEEE Transactions on Image Processing*, 31:1084–1096, 2022. 1

[46] Jue Wang and Lorenzo Torresani. Deformable video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14053–14062, June 2022. 2

[47] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1

[48] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1078–1084. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 2

[49] ByungIn Yoo, Youngjun Kwak, Youngsung Kim, Changkyu Choi, and Junmo Kim. Deep facial age estimation using conditional multitask learning with weak label expansion. *IEEE Signal Processing Letters*, 25(6):808–812, 2018. 1

[50] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3ae: Exploring the limits of compact model for age estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12587–12596, 2019. 6

[51] Ke Zhang, Na Liu, Xingfang Yuan, Xinyao Guo, Ce Gao, Zhenbing Zhao, and Zhanyu Ma. Fine-grained age estimation in the wild with attention lstm networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):3140–3152, 2019. 1

[52] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 4, 5

[53] Yunxuan Zhang, Li Liu, Cheng Li, et al. Quantifying facial age by posterior of age comparisons. *arXiv preprint arXiv:1708.09687*, 2017. 1

[54] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person reidentification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019. 3

[55] Yu Zhu, Yan Li, Guowang Mu, and Guodong Guo. A study on apparent age estimation. In *proceedings of the IEEE international conference on computer vision workshops*, pages 25–31, 2015. 1